


CIVIC

DIGITAL

FELLOWSHIP




Fuzzy Address Matching of 1990s-Present Census Data

Anthony Xiang

Supervised by Katie Genadek and Christian Moscardi

Economic Reimbursable Surveys Division

Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. 
All results were approved for release by the U.S. Census Bureau, authorization number CBDRB-FY20-ERD002-033.

United States[®]
Census
2020

Overview

- The 1990 decennial census was previously **unlinkable**.
- By **linking datasets on address**, we can enable longitudinal researchers to work across **decades of data** (1990-present).
- **Goal:** create id to id crosswalks based on addresses for:
 - 1993 NSCG – 1990 Address Control File (ACF)
 - 1990 ACF – 1990 Decennial Census
 - 1990 ACF – Master Address File (2000s and onward)



Challenges

- Address entries are not consistent

Ex: 1600 East Pennsylvania Avenue vs. 1600 Pensylvania Av E

Dataset	Size
National Survey of College Graduates (NSCG)	150,000 records
1990 Decennial Census	102 million records
Address Control File (ACF)	104 million records
Master Address File (MAF)	106 million records

All results were approved for release by the U.S. Census Bureau, authorization number CBDRB-FY20-ERD002-033.

- MAMBA software: would take weeks to months
- MAFMatch: Backlogged
- Google Maps API: \$400,000+

TF-IDF

- Trigrams/TF-IDF/KNN algorithm
 - Break up address by letters
 - Numbers used as similarity score
- TF-IDF distance as an accuracy metric
- Threshold set at distance = 1
- Hierarchical blocking

13 hicks street

	distance	address_A	address_B	id_A	id_B
4144	0.90	13 hicks street	13 hick ave	2181	2181
4444	0.75	5 bellinger circuit	5 bellinger crt	4076	4076
1317	0.00	19 sheaffe street	19 sheaffe street	1274	1274
335	0.69	3 wicks road	3 wicks rquad	2735	2735
3149	0.00	6 majura road	6 majura road	3098	3098
3144	0.68	7 langdon avenue	7 langdonavenue	236	236

Addresses in table are dummy data from a Python package, Python Record Linkage Toolkit

Results and Method Comparison

- 1993 NSCG – 1990 ACF -> **99.74%** successful match rate¹
- 1990 ACF – 1990 Decennial Census -> **96.96%** match rate
- 1990 ACF – Master Address File (2000s and onward) -> **83.31%** successful match rate^{2,3}
- MAMBA software: would take weeks to months
- MAFMatch: Backlogged
- Google Maps API: \$400,000+
- TF-IDF: **few hours and free software!**

1 Distance threshold < 1, name match, or similar address match

2 Distance threshold < 0.8

3 All results were approved for release by the U.S. Census Bureau, authorization number CBDRB-FY20-ERD002-033.

TF-IDF Open Sourced within Census

```
import tfidf_wrapper.tfidf_wrapper as lh
linked_blocked3 = lh.linkDatasets(dfA, dfB, 'address_A', 'address_B', 'rec_id', 'rec_id', block="state")
```

Many benefits and applications!

- Fast, easy, and accurate solution to record linkage problems
- Link datasets in only **two lines of code**
- Useful for names, businesses, addresses, etc.

<https://vc1.csvd.census.gov/xiang001/tfidf-record-matching>

Addresses in table below are dummy data from a Python package, Python Record Linkage Toolkit

	distance	address_A	address_B	id_A	id_B
4144	0.90	13 hicks street	13 hick ave	2181	2181
4444	0.75	5 bellinger circuit	5 bellinger crt	4076	4076
1317	0.00	19 sheaffe street	19 sheaffe street	1274	1274
335	0.69	3 wicks road	3 wicks rquad	2735	2735
3149	0.00	6 majura road	6 majura road	3098	3098
3144	0.68	7 langdon avenue	7 langdonavenue	236	236

Thank you!